Published in final edited form as:
Cancer Epidemiol Biomarkers Prev. 2020 April ; 29(4): 777–786. doi:10.1158/1055-9965.EPI-19-0842.

Insights from Adopting a Data Commons Approach for Large-Scale Observational Cohort Studies: The California Teachers Study

James V. Lacey Jr.[1,*], **Nadia T. Chung**[1], **Paul Hughes**[2], **Jennifer L. Benbow**[1], **Christine Duffy**[3], **Kristen E. Savage**[1], **Emma S. Spielfogel**[1], **Sophia S. Wang**[1], **M. Elena Martinez**[4], **Sandeep Chandra**[2]

[1]Department of Computational and Quantitative Medicine, City of Hope, Duarte, CA.

[2]Sherlock, San Diego Supercomputer Center, University of California, San Diego, San Diego, CA.

[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA.

[4]Department of Family Medicine and Public Health, University of California, San Diego, San Diego, CA.

Abstract

Background: Large-scale cancer epidemiology cohorts (CECs) have successfully collected, analyzed, and shared patient-reported data for years. CECs increasingly need to make their data more Findable, Accessible, Interoperable, and Reusable, or FAIR. How CECs should approach this transformation is unclear.

Methods: The California Teachers Study (CTS) is an observational CEC of 133,477 participants followed since 1995–1996. In 2014, we began updating our data storage, management, analysis, and sharing strategy. With the San Diego Supercomputer Center, we deployed a new infrastructure based on a Data Warehouse, to integrate and manage data; and a secure and shared workspace with documentation, software, and analytic tools that facilitate collaboration and accelerate analyses.

Results: Our new CTS infrastructure includes a Data Warehouse and data marts, which are focused subsets from the Data Warehouse designed for efficiency. The secure CTS workspace utilizes a Remote Desktop service that operates within a HIPAA and FISMA compliant platform. Our infrastructure offers broad access to CTS data; includes statistical analysis and data visualization software and tools; flexibly manages other key data activities (e.g., cleaning, updates, & data sharing); and will continue to evolve to advance FAIR principles.

Conclusion: Our scalable infrastructure provides the security, authorization, data model, metadata, and analytic tools needed to manage, share, and analyze CTS data in ways that are consistent with the NCI's Cancer Research Data Commons Framework.

[*]Correspondence: James V. Lacey, Jr.; Department of Computational and Quantitative Medicine; City of Hope, 1500 East Duarte Road, Duarte, CA 91010-3000. Telephone: 626-218-3837; Fax: 626-471-7308; jlacey@coh.org.

Conflict of Interest: The authors declare no potential conflicts of interest.

Impact: The CTS's implementation of new infrastructure in an ongoing CEC demonstrates how population sciences can explore and embrace new cloud-based and analytics infrastructure to accelerate cancer research and translation.

Keywords

data warehouse; cloud computing; cancer epidemiology; data science; big data

INTRODUCTION

Cancer epidemiology cohorts (CECs) are an essential component of the cancer research enterprise (1). CECs that enroll large numbers of participants, collect detailed data on multiple exposures and outcomes, and follow those volunteers over many years provide a unique source of real-world data for generating and testing hypotheses. The size, scale, and duration of CECs require substantial investments of time, financial resources, and experienced personnel (2), but CECs directly facilitate etiologic and translational research on environmental, lifestyle, genomics, and social factors that affect cancer risks, survivorship, and disparities (3).

As in other areas of biomedical research, data are the lifeblood of a CEC. Across all aspects of biomedical research, data and data science are changing faster than ever (4). New developments in information technology and data science present special challenges for CECs. Information technology is the "single most important infrastructure component for ensuring high data quality and cost control in large studies" (2), but managing a CEC's data and IT infrastructure—or changing a CEC's IT infrastructure during follow-up—can have high financial and opportunity costs. A 2013 publication in CEBP about transforming cancer epidemiology for the 21st century recommended that cancer epidemiology "develop, evaluate, and use novel technologies appropriately" (5). In 2013, the California Teachers Study (CTS), a multi-site prospective CEC of 133,479 women, found itself using the same data strategy, analysis infrastructure, and IT resources that it had used since 1995, when the CTS began. The CTS stored its data on local network drives at CTS investigators' institutions—but this created data silos that hindered real-time collaboration. The CTS manually merged, updated, and distributed individual and summary datasets for analyses and data sharing—but this was time-consuming (2) and less efficient than other CECs that were storing data in a centralized server (6) or repository (7) environments. In 2014, the CTS adopted a new cloud-based approach for storing, analyzing, and sharing CTS data in one common, secure, shared environment. This manuscript describes the development, deployment, and features of this infrastructure; and we conclude by discussing strengths and limitations of our approach.

MATERIALS AND METHODS

BACKGROUND: THE CALIFORNIA TEACHERS STUDY

The CTS began in 1995–1996, when 133,479 active or recently retired public school professionals completed a self-administered survey and consented to be prospectively followed and have their data used for cancer and women's health research. Most of the

CTS participants lived in California at the time of enrollment. Except for two participants who withdrew during follow-up, all participants have been followed continuously since. Characteristics of CTS participants at baseline have been described (8). Five follow-up surveys (1997–1998; 2000–2001; 2005–2006; 2012–2014; and 2017-present) collected additional self-reported data. The CTS website includes details and copies of these surveys (9).

Cancers and other endpoints are obtained via numerous linkages. We annually link CTS data to the California Cancer Registry (CCR) to identify new cancer diagnoses, with detailed tumor data, for CTS participants. Probabilistic linkage software allows multiple-pass matching to maximize the number of true matches, minimize the amount of manual review, and avoid errors due to crucial blocking variables. The California Office of Statewide Health Planning and Development (OSHPD) collects information on up to 25 diagnoses and up to 21 procedures per admission for each non-Veterans-Administration hospitalization in California. The CTS has conducted annual probabilistic record linkage with OSHPD to enhance follow-up, investigate comorbidities, and identify non-cancer disease outcomes. Annual linkages with California Department of Public Health (CDPH) mortality files, the national Social Security Administration (SSA) Death Master File, and National Death Index (NDI) data provide dates and causes of death for CTS participants who died.

Before annual CTS mailings, the CTS members’ address data files are processed electronically by a U.S. Postal Service-designated service agency to identify recent changes. We continuously collect name and address changes from the CTS website, toll-free phone number, participants’ emails, questionnaires, and pre-paid postcards in mailed annual newsletters. Additional information about CTS publications, data, and biospecimens is available at www.calteachersstudy.org.

By 2014, four sites actively managed portions of the CTS data. The University of Southern California (USC) managed cancer, mortality, hospitalization, and address linkages; maintained the participants’ identifying information; and stored some biospecimens. The University of California, Irvine (UCI) stored other biospecimens. The Cancer Prevention Institute of California retained dietary, environmental, and some residential data and stored some biospecimens. City of Hope (COH) managed an ongoing large-scale biospecimens project (10), hospital records, and all analytic datasets. COH updated CTS data after data cleaning, new surveys, or linkages; sent replacement datasets to each site; and made and sent custom datasets for each new analysis or data-sharing project.

CONSIDERATION OF POTENTIAL NEW TECHNOLOGIES

In 2014, the CTS identified three areas for potential improvement. These became the essential features that the CTS sought in its new data strategy. First, we needed to eliminate the data silos associated with having CTS data at each site. Second, to facilitate collaboration and data sharing, we needed a flexible workspace in which all CTS data, analyses, and results would be transparent and available to all collaborators. Third, we wanted to accelerate the processes of updating, replacing, merging, and distributing data to investigators by replacing the manual processes that the CTS had used with automated activities that occurred directly on the data.

Storing all CTS datasets and code in a centralized library or web portal would eliminate data silos and create a shared workspace, but too many data steps—e.g., updating data, creating analysis-specific datasets, sharing data with external investigators—would still be manual, time-consuming, and difficult to scale. Our 2013–2016 biobanking project (10) introduced us to cloud-computing platforms that went beyond data storage by directly integrating and presenting data in user-friendly and efficient formats. After considering different ways to achieve all three of the objectives described above, we concluded that the CTS needed a Data Warehouse (DW). In 2015, the CTS received grant award U01-CA199277, whose primary objective was to democratize all CTS data through a new and more efficient DW-based core infrastructure.

A CTS DATA WAREHOUSE

DWs have existed for over 50 years and have enabled numerous industries to convert data into information and insight that improves outcomes (11). DWs integrate, conform, and store data from multiple sources and then present those data in specific ways that directly facilitate data mining, research, analytics, and reports—all while providing full data lineage. Data warehouses are more than just databases; they typically exist at an enterprise or organizational level to integrate large amounts of diverse data from different sources for business intelligence. Large CECs, such as the CTS, are not businesses but can resemble an “enterprise” or small organization. A dedicated CTS DW would meet all of our current and future storage and analytic needs (Table 1).

The CTS chose the San Diego Supercomputer Center (SDSC), which is part of the University of California, San Diego (UCSD) to build, manage, and support the CTS DW. Established in 1985 as 1 of 4 original National Science Foundation supercomputer centers, SDSC provides innovative computing, cyberinfrastructure, services, and expertise. SDSC includes Sherlock Cloud, a HIPAA-compliant cloud platform designed specifically to meet the security, regulatory, and compliance needs of users—such as the CTS—who manage personal health information (PHI). Sherlock recently became one of the first academic institutions to offer secure and compliant managed services through a hybrid cloud [i.e., one that uses both private (SDSC) and public (Amazon Web Services)] and was selected to deploy a HIPAA-compliant data management system (DMS) for managing healthcare data across the entire University of California system (12).

DESIGNING AND BUILDING THE DATA WAREHOUSE.

Building the DW followed a six-step process shown in Figure 1. First, the CTS team identified the CTS data that the DW would contain: all CTS exposure data from the baseline and follow-up surveys, including derived and calculated variables (e.g., body mass index); all cancer, hospitalization, and mortality outcomes; biospecimen data; participant residential history data; and participant-level contact information (e.g., first and last name for linkages). The CTS formatted its datasets and accompanying documentation in standard ways to populate the data tables that would provide the foundation of the DW.

Next, we chose a Microsoft Remote Desktop Connection (RDC, i.e., remote desktop protocol) user interface operating within a secure environment. Authorized users launch

RDC to access the new CTS data environment; after entering their username and password, a Windows desktop appears with file directories, software, and CTS resources.

Third, we collocated copies of CTS data within the new CTS environment. This enabled us to replicate sample analyses within the environment and verify that the data we had loaded into the environment produced results identical to analysis of the CTS's previous, pre-DW data.

Fourth, the SDSC team built a hybrid DW with an Enterprise scoped core and subject-specific data marts (e.g., a cancer data mart, a biospecimen data mart, etc.), using a set of industry standards and best practices. Data marts are more narrowly focused extracts of specific types of data from the DW that are designed for performance, analysis, and reporting. Based on the CTS data and analytic needs, SDSC combined models from Kimball (13), Inmon (14), and Entity Attribute Valued (15). SDSC used a traditional DW architecture with Windows 2012 as the foundation operating system across three servers: a relational database server, an ETL (extract, transform, and load) server, and a file server. The relational database platform uses a Microsoft 2012 SQL server with transparent data encryption. Informatica 9.6.1 provides a mature, best-of-breed ETL server, and Toad data modeler provides the Modeling toolkit. This architecture—i.e., the combination of the data model and the three servers—differentiates the CTS DW from a centralized data enclave or repository.

To convert CTS data into the DW and Data Marts, we provided source datasets that SDSC loaded into its secure, collocated staging area. SDSC organized those CTS data into a master database based on a participant-centered data model that facilitates CEC analyses of individual participants over time. The ETL server extracts and normalizes, in a third normal form (3nf), all CTS data about participants into a structure that facilitates CEC reporting and analysis. From there, a dimensional star schema model enables Marts that present CTS data from the DW to researchers for exploration, reporting, and statistical analysis. By September 2016, the DW was complete and ready to support CTS analyses. Although SDSC offers public and hybrid cloud options, the CTS DW currently utilizes SDSC's private cloud services.

RESULTS

For simplicity, "the environment" refers to the CTS DW, data marts, and shared workspace accessible via the secure RDC.

Data Security

The CTS and SDSC jointly manage a multi-level data security model. The DW contains all CTS analytic data, plus PHI for study management (e.g., residential addresses for geospatial research). Users can request access through the CTS website (see www.calteachersstudy.org/for-researchers). The CTS approves requests and assigns a user role (see below). The SDSC team creates accounts and user names. The CTS then sends the RDP program, credentials, and a detailed User Guide to new users. SDSC maintains an Active Directory of all authorized users. Users are assigned one of three roles that determine the type and amount of information they can access within the environment. All three roles are limited to read-only access within the DW database.

The Manager role is limited to a small number of CTS personnel who can access all data and fields in all databases, including historical tables. Managers work with SDSC to manage the data and environment. The Analyst role includes CTS team members and most external collaborators. Analysts can access everything except staged data and sensitive/PHI data and can use the secure managed file transfer (MFT) protocol for moving files into and out of the environment. The Researcher role is “read-only”: researchers can view data in Data Marts but cannot modify data or use MFT.

When users attempt to login to the CTS data environment using the RDC, the Active Directory provides account authentication. A user’s role determines what objects they can view, permissions they have, and actions they can perform. Views over the base data tables provide an additional security layer: base tables contain the actual normalized data, and views are virtual tables whose contents are defined by a query that arises from the actions that user takes—verified by that user’s role—within the environment. These views allow us to hide sensitive data; mask or filter data as needed; hide data complexity by modeling complex data joins; and rename base table field names to more commonly understood terms.

Remote Desktop Environment

The environment brings users and software to the data and allows SDSC and CTS to manage and audit the environment. The Remote Desktop includes file folders, office software (MS Word, Excel, and Adobe), and analytic software (e.g., SAS, Cary, NC, and R). Additional software or tools can be added at any time. To add software, the CTS reviews and approves requests and, if necessary, purchases licenses. SDSC performs a standard security review and then manages the installation and ongoing support. Shared folders include documentation, background materials, data dictionaries, and preformatted and standardized CTS cohort data for open-ended, no-approval-needed exploration. Useful administrative documents (e.g., IRB approval) are there. To encourage use of the shared workspace, each project receives a specific network folder for its team to use, but each approved user also receives a “My Documents” folder for their personal use. Everyone who accesses the environment works in the same space and accesses the same data and software as all other authorized users, but everyone does so as individual users, under their own user account.

Data Management

We agreed on and implemented a a common CTS file structure, naming convention, and data key for linking data across all CTS data sources. All data are mapped to CTS data domains and subject areas. Using the secure managed file transfer (MFT) process, CTS data managers upload data (e.g., new CTS data or any external data, such as publicly available geospatial data, needed for projects) into the dedicated staging area. All file transfers trigger standard notifications, alerts, and footprints as part of Sherlock’s track, trace, and resolve functionality. After upload, all files follow a standard “Inbound, Validate, Loading, and Archive” onboarding cycle. Inbound files are validated against expected columns and rows as defined by each upload's standard data dictionary and definitions files. Files that fail any of these checks move to the "Invalid" area, and a report is generated. Files that fail to load move to the "Failed" area. These areas allow those files to be explored and fixed before they overwrite any existing data. A file directory scanner reads all files, manages their status, and generates the required metadata for each. A second process reviews the results of the file directory scanner, identifies data to load, and then either matches those data to existing data or performs any required data conversions. A final rule-based engine manages, maintains, and executes all of the stages of the loading cycle. Some data (e.g., updated versions of linkage data) are managed through common sets of rules with significant automation. If data require additional processing or conversions, manual steps can be added to ensure integration, consistency, and quality. As soon as new data complete this cycle, those data are available, fully integrated at the participant level with all other CTS data, to authorized users. To ensure version control, the environment provides complete data lineage at all data levels, from source to column to observation. When new data replace existing data, the existing data are expired but not deleted. This preserves the ability to retrieve any previous or expired CTS data at any future time point.

Our use of Informatica's PowerCenter as the ETL tool provides additional version control. This tool provides built-in runtime metadata on workflows and sessions, including source/target, success/failure row counts, timings, throughput and errors. We also use a versioned Repository option that 1) enforces a check-out, check-in versioning policy applied to individual objects (e.g., source or target tables and transformations) and to collection of objects (e.g., mappings, sessions, and workflows); 2) provides a full history of version changes, to allow comparisons and point-in-time recovery; and 3) extends into the deployment capability, providing history of deployments, rollback capability and comparisons. These repository backups and object-level exports to XML files can also allow integration to other source control management systems and backup\restore strategies, if required. Together with the data lineage, these versioning and runtime statistics provide a mature software development life cycle capability.

Data Marts.

The DW includes all of the integrated and normalized CTS data. However, asking users to use the large, complete, and comprehensive data for every CTS query or analysis is inefficient. The SDSC therefore created six Data Marts to facilitate common CTS queries and analyses. A Cancer Mart combines all of the data on cancer outcomes among CTS participants to enable users to explore detailed cancer endpoints. A Life Event Mart integrates participant-level data on key life events, such as dates of birth, death, entry into the CTS, and diagnosis of first cancer, to standardize follow-up metadata and give every project the flexibility it needs. A Questionnaire Mart integrates all of the CTS questionnaire responses at the participant level. Similar Marts for CTS biospecimens and for hospitalization events facilitate exploration of those specific CTS data by combining and structuring the data from the CTS DW in ways that help users accomplish their CTS goals. A Geospatial Research Mart includes extensive geographical reference data (e.g., U.S. Census tract data) to facilitate efficient integration of any geospatial data into the CTS DW for individual-level CTS analysis. Marts are flexible and expandable and can be joined (e.g., biospecimens and cancer) for exploration and analysis.

Data Analyses.

The primary use of CTS data is for testing hypotheses using specific data subsets. Our infrastructure standardizes the processes of selecting cohorts, applying inclusion and exclusion criteria, and requesting specific data. After the CTS approves a research proposal (16), the investigator completes a menu-based, sequential question-and-answer template to capture and document all of her detailed analytic decisions and request those analytic data. We transfer those choices into a SQL-based template that extracts, transforms (if needed), and presents those specific data from the Mart(s) to that user for analysis within the Environment. This SQL template eliminates the need for every investigator to merge data, apply exclusions, check for duplicates, or manage data. We give investigators a 1-line SAS program that runs their SQL template; they can run, edit, and expand these programs to analyze their specific data. The SQL template can be modified, either by us or the investigator, whenever their analysis changes. Our goal is to leverage the core data model and Mart scheme to efficiently deliver specific data to users, so that they can spend their time analyzing, rather than managing, data.

Every project gets a dedicated folder in the shared workspace. Each folder includes three items: 1) copies of their SQL and 1-line SAS codes, as described above; 2) a custom data dictionary for that project; and 3) a data visualization workbook (Tableau software (Seattle, WA)) that provides a standardized list of all of the sequential exclusions that were applied for that analysis, and that includes ready-to-go, interactive visualization workbooks that display the analytic data for point-and-click queries. Figure 2 provides examples of CTS data visualizations.

Data Sharing.

Before 2015, the CTS prepared and sent physical datasets to data-sharing recipients. The CTS now shares data in three ways. Data-sharing recipients can analyze their CTS data within the shared workspace of the CTS data environment. They can also start their analysis in the CTS environment (e.g., review data or covariates) and then ask the CTS to send a copy of those data from the DM to another site, such as a data coordinating center. The Sherlock cloud also supports an Application Programming Interface, or API, for virtual sharing: CTS data could be securely exposed for analysis in an external or virtual environment.

Table 2 shows the core services required of a NCI Cancer Research Data Commons (CDRC) (17) and how the CTS meets those requirements. As of July 2019, the environment includes 1 terabyte of storage and can expand as needed. The core currently includes 46 entities, 866 attributes, 37 relationships, 51 keys, and 87 views. The mart model includes 21 entities, 367 attributes, 29 relationships, 34 indexes, and 21 keys. The DW includes over 160 Million data cells from CTS surveys; 500,000 hospitalization records; and cancer data for over 32,000 cancers diagnosed during CTS follow-up.

DISCUSSION

Since 2014, the CTS replaced its data storage, management, and analytic infrastructure with a centralized, integrated, and automated approach. The CTS now offers all users a shared and secure workspace with common data, documentation, software, and analytic tools to facilitate use and data sharing.

The CTS is not the first or only CEC to adopt a bring-users-to-the-data approach. The Women's Health Initiative (WHI) uses a virtual data enclave (VDE) for its datasets and analysis (7). The NCI's Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial datasets are available through the web-based Cancer Data Access System (18). The Nurses' Health Studies provides documentation, analysis tools, and analysis examples and require analyses to occur on the UNIX server that stores their data (6). Although those approaches would have improved the CTS by centralizing data and analysis, we instead implemented a full DW that provided three benefits for the CTS. First, the DW's data model, normalized data, and combination of data, file, and ETL servers enable us to develop reports, Marts, workflows, and visualizations that directly facilitate epidemiologic queries and analyses. Second, the DW includes a framework for assigning quality, completeness, and trustworthiness rules to the integrated data. The CTS previously did this within individual analyses (i.e., via SAS code). Transplanting those CEC-specific nuances from downstream statistical analyses, which are difficult to scale (19), to the upstream core data increases transparency and sustainability. Third, the DW includes the type of extensive metadata that tend to be underutilized (4) but provide valuable insights into how and what CTS data are used.

Different CECs have different needs, and data lakes (20) and late-binding data warehouses (21) could support CECs' data needs. No two CECs are identical, and no single data analytics platform or cloud-computing strategy will work for every CEC. The National Cancer Data Ecosystem provides a roadmap that population sciences can use to make their data more FAIR—Findable, Accessible, Interoperable, and Reusable. How CECs do so—e.g., via a DW, data lake, open-source notebook, or data enclave, etc.—matters less than whether their infrastructure provides all of the required core services and interoperability (4).

We partnered with SDSC Sherlock because our strategy required skills and expertise we lacked. Cloud-based storage, compute, and software as a service are now commodities that many providers offer. SDSC provided managed services expertise to help configure a CEC-specific environment and user experience. The SDSC team understood our analytic needs and suggested potential solutions. Although the NCI CRDC Framework emerged after we began building and deploying our DW, our infrastructure meets the CRDC's requirements (22, 23). This demonstrates that the DW infrastructure SDSC offers is consistent with the research community's current standards and reaffirms that DWs have the flexibility and scalability to support CEC research.

Our infrastructure has numerous strengths. The environment increases access to CTS data, removes many barriers to use, and accommodates CEC data storage, cleaning, updating, analyzing, and sharing. Universal access to common CTS resources and documentation helps launch new collaborations. The multi-level security framework provides strong data protections. We improved version control and eliminated concerns about outdated data. The environment includes data visualization to facilitate reporting, exploration, and utilization of multi-dimensional CTS data. This transition prompted us to standardize and update our proposal review process (16), helped us identify data and process flaws, and led to improved CTS data management and governance. The CTS DW's metadata model helps us track when and how specific CTS data are used. These metadata could help catalyze the overdue development of a metadata model (4) for CEC data within the cancer research enterprise.

Our infrastructure also has limitations. In 2014, no template existed for a CEC-specific DW. We chose a custom data model to hedge against potential poor fit of CTS data to other observational data models based on electronic medical records (EMRs). Fitting the CTS to a fully open data model, such as OHDSI, would further increase interoperability. As an early adopter among CECs, building our DW required significant up-front investment to develop a data model, configure the user interface, and convert 20 years' worth of existing CTS datasets into a single integrated DW environment. Replicating this process in another CEC would require less investment today because increasingly more options and examples exist for transitioning to fully FAIR data. Our CTS DW team embraced the opportunity to learn new data science (4) and data warehousing skills. Pivoting from every CTS investigator analyzing her own copy of CTS data to all investigators using shared resources requires a conceptual shift in focus from the individual investigator to the broader user community. There were differing levels of uptake across the CTS. The Windows remote desktop provides a secure environment with a familiar user interface, but this approach also puts limits on external connections. Other choices, such as workbenches, also balance security and interoperability. The growth of healthcare research DWs is generating a deeper understanding of the organizational, behavioral, and technical challenges of these types of transitions (24–27).

Our CTS infrastructure will continue to grow and evolve. In the next year, we will expand the geospatial, comorbidities, and biospecimen tools for query and analyses. The process of exploring, requesting, and revising project-specific data can be further automated. Development of an Application Programming Interface (API) can facilitate data collection and sharing.

CECs capture, store, and analyze real-world data that is increasingly important in cancer research (4). Moving beyond data storage and having more high-value CEC data available through the NCI CRDC (4) could generate even greater return on the significant investments made and resources that exist in CECs today (19). The CEC community has a strong history of data sharing and agrees with the need to continue sharing CEC data consistent with FAIR principles (28,29). There are also reasonable questions about how best to replicate the individuality and complexity of CEC-based analyses in a CRDC environment (30). Compliance with FAIR principles is not only "either/or" but instead exists on a spectrum. Participation by CECs in the NCI CRDC need not be "all or nothing" but could instead occur in phases that help the CEC community refine the specific tools and approaches that maximize benefits for all stakeholders.

Our CTS DW and shared workspace is consistent with the NCI CRDC Framework and provides comprehensive yet flexible infrastructure for storing, managing, analyzing, and sharing CEC data. Rather than being a static endpoint in itself, the new CTS data infrastructure provides a foundation for continuous and ongoing development and expansion. Like every NCI-funded CEC, the CTS has unique features but shares enough common characteristics to facilitate widespread data harmonization, pooling, and sharing. Our transition of CTS core data infrastructure affirms that a DW, and the overall NCI CRDC Framework, can support essential CEC activities. The scalability and efficiencies of the CTS approach represents a model other new and existing CECs could adapt and use for their infrastructure.

ACKNOWLEDGEMENTS

The California Teachers Study and the research reported in this publication were supported by the National Cancer Institute of the National Institutes of Health under award number U01-CA199277; P30-CA033572; P30-CA023100; UM1-CA164917; and R01-CA077398. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

The collection of cancer incidence data used in the California Teachers Study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute. The opinions, findings, and conclusions expressed herein are those of the author(s) and do not necessarily reflect the official views of the State of California, Department of Public Health, the National Cancer Institute, the National Institutes of Health, the Centers for Disease Control and Prevention or their Contractors and Subcontractors, or the Regents of the University of California, or any of its programs.

The authors would like to thank the California Teachers Study Steering Committee that is responsible for the formation and maintenance of the Study within which this research was conducted. A full list of California Teachers Study team members is available at https://www.calteachersstudy.org/team."

All of the data associated with this publication and in the California Teachers Study are available for research use. The California Teachers Study welcomes all such inquiries and encourages individuals to visit https://www.calteachersstudy.org/for-researchers.

References

1. Core Infrastructure and Methodological Research for Cancer Epidemiology Cohorts (U01). 7/16/2019; Funding Opportunity Announcement (FOA) Number PAR-17–233]. Available from: https://grants.nih.gov/grants/guide/pa-files/PAR-17-233.html
2. Manolio TA, Weis BK, Cowie CC, Hoover RN, Hudson K, Kramer BS, et al. New models for large prospective studies: is there a better way? Am J Epidemiol. 2012;175:859–66. [PubMed: 22411865]
3. Colditz GA, Winn DM. Criteria for the evaluation of large cohort studies: an application to the nurses' health study. J Natl Cancer Inst. 2008;100:918–25. [PubMed: 18577745]
4. NCAB Ad Hoc Working Group on Data Science. Data Science Opportunities for the National Cancer Institute. Report of the National Cancer Advisory Board Working Group on Data Science. 7/16/2019. Available from: https://deainfo.nci.nih.gov/advisory/ncab/workgroup/DataScienceWG/WGJune2019recommendations.pdf.
5. Khoury MJ, Lam TK, Ioannidis JP, Hartge P, Spitz MR, Buring JE, et al. Transforming epidemiology for 21st century medicine and public health. Cancer Epidemiol Biomarkers Prev. 2013;22:508–16. [PubMed: 23462917]
6. Nurses' Health Study. For researchers. 7/16/2019; Available from: https://www.nurseshealthstudy.org/researchers

7. Women’s Health Initiative. WHI Virtual Data Enclave (VDE). 7/16/2019; Available from: https://www.whi.org/researchers/data/SitePages/VDE.aspx

8. Bernstein L, Allen M, Anton-Culver H, Deapen D, Horn-Ross PL, Peel D, et al. High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). Cancer Causes & Control. 2002;13:625–35. [PubMed: 12296510]

9. California Teachers Study. March 27, 2019; Available from: www.calteachersstudy.org

10. Faster, Safer, Cheaper, Better: How CRM and cloud computing can help studies collect, store, use, and share data. NCI CBIIT Speaker Series 2014. 7/16/2019. Available from: https://youtu.be/BezLC1-jwkQ

11. Sanders D Wal-Mart and the Birth of the Data Warehouse. 2013. 7/16/2019; Available from: https://www.healthcatalyst.com/wal-mart-birth-of-data-warehouse/

12. Sherlock at Work. 7/16/2019; Available from: https://sherlock.sdsc.edu/about/

13. The Kimball Group. Kimball Techniques. 7/16/2019; Available from: https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/

14. Kimball vs. Inmon Data Warehouse Architectures. 7/16/2019; Available from: https://www.zentut.com/data-warehouse/kimball-and-inmon-data-warehouse-architectures/

15. Wikipedia. Entity-attribute-value model. 7/16/2019; Available from: https://en.wikipedia.org/wiki/Entity%E2%80%93attribute%E2%80%93value_model

16. The California Teachers Study. For Researchers. March 25, 2019; Available from: https://www.calteachersstudy.org/for-researchers

17. NCI Cancer Research Data Commons. 7/16/2019; Available from: https://datascience.cancer.gov/data-commons

18. PLCO Cancer Data Access System. 7/16/2019; Available from: https://biometry.nci.nih.gov/cdas/plco/

19. Ad Hoc Working Group on Strategic Approaches and Opportunities in Population Science E, and Disparities. Report on National Cancer Institute (NCI) Extramural Cancer Epidemiology Cohort Studies. 7/16/2019; Available from: https://deainfo.nci.nih.gov/advisory/ncab/workgroup/StrategicApproaches/WGonPopSciEpiDisparities-FinalReport.pdf

20. What is a data lake? 7/16/2019; Available from: https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

21. Yang E, Scheff JD, Shen SC, Farnum MA, Sefton J, Lobanov VS, et al. A late-binding, distributed, NoSQL warehouse for integrating patient data from clinical trials. Database. 2019 Jan 1. pii: baz032. 10.1093/database/baz032. [PubMed: 30854563]

22. Grossman RL. Progress Toward Cancer Data Ecosystems. Cancer J. 2018;24:126–30. [PubMed: 29794537]

23. Grossman RL. Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data. Trends Genet. 2019;35:223–34. [PubMed: 30691868]

24. Baghal A, Zozus M, Baghal A, Al-Shukri S, Prior F. Factors Associated with Increased Adoption of a Research Data Warehouse. Stud Health Technol Inform. 2019;257:31–5. [PubMed: 30741168]

25. Doria-Rose VP, Greenlee RT, Buist DSM, Miglioretti DL, Corley DA, Brown JS, et al. Collaborating on Data, Science, and Infrastructure: The 20-Year Journey of the Cancer Research Network. EGEMS. 2019;7(1):7. [PubMed: 30972356]

26. Karami M, Rahimi A, Shahmirzadi AH. Clinical Data Warehouse: An Effective Tool to Create Intelligence in Disease Management. Health Care Manag. 2017;36:380–4.

27. Seneviratne MG, Seto T, Blayney DW, Brooks JD, Hernandez-Boussard T. Architecture and Implementation of a Clinical Research Data Warehouse for Prostate Cancer. EGEMS. 2018;6(1):13. [PubMed: 30094285]

28. Swerdlow AJ, Harvey CE, Milne RL, Pottinger CA, Vachon CM, Wilkens LR, Gapstur SM, Johansson M, Weiderpass E, Winn DM. The National Cancer Institute Cohort Consortium: an international pooling collaboration of 58 cohorts from 20 countries. Cancer Epidemiol Biomarkers Prev. 2018 Jul 17. pii: cebp.0182.2018.

29. NCI Cohort Consortium Strategic Plan. 12/2/2019; Available from: https://epi.grants.cancer.gov/Consortia/nci-cohort-consortium-strategic-initiatives-july19-2019.pdf

30. NIH Data Management and Sharing Activities Related to Public Access and Open Science. 12/2/2019; Available from: https://osp.od.nih.gov/scientific-sharing/nih-data-management-and-sharing-activities-related-to-public-access-and-open-science/

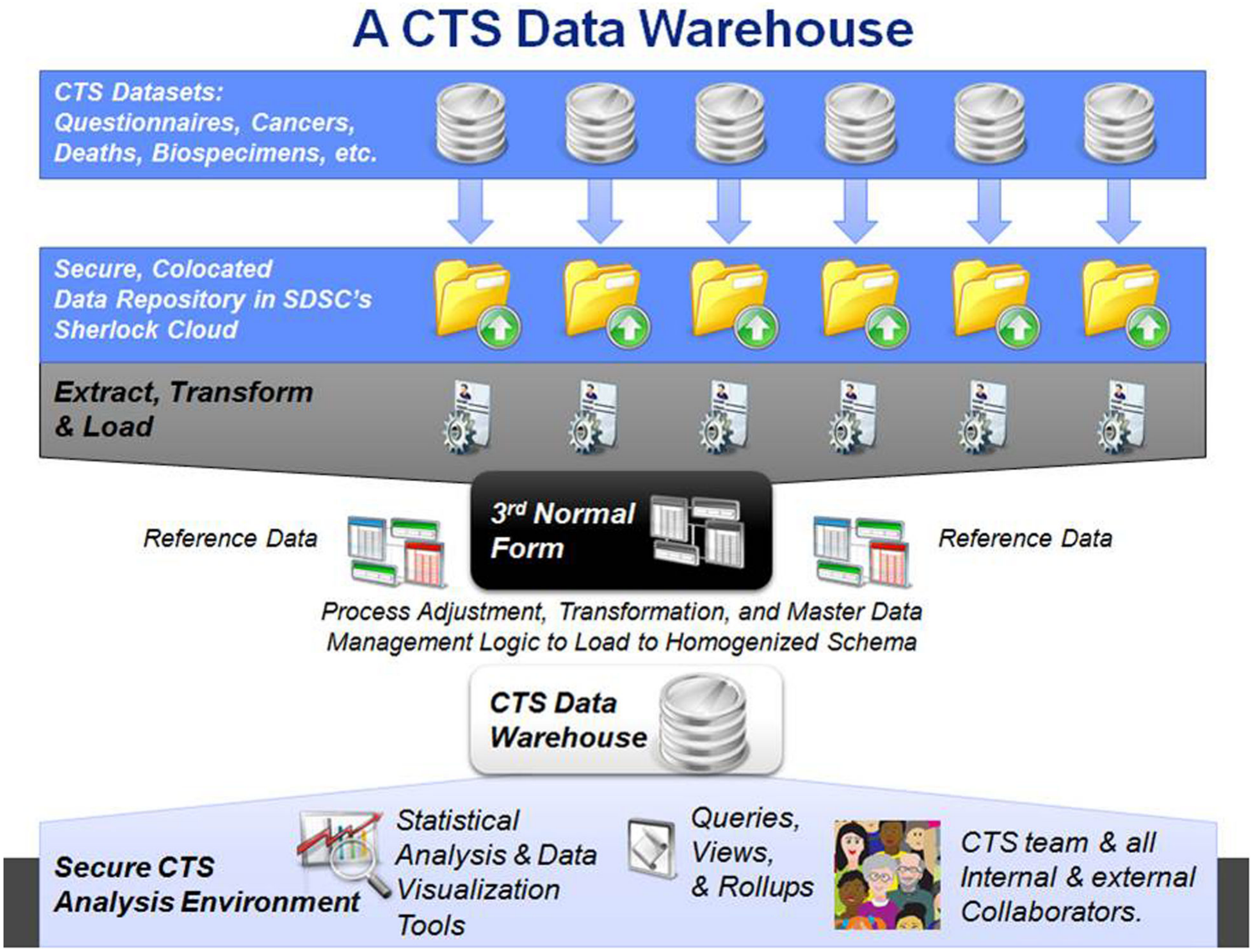
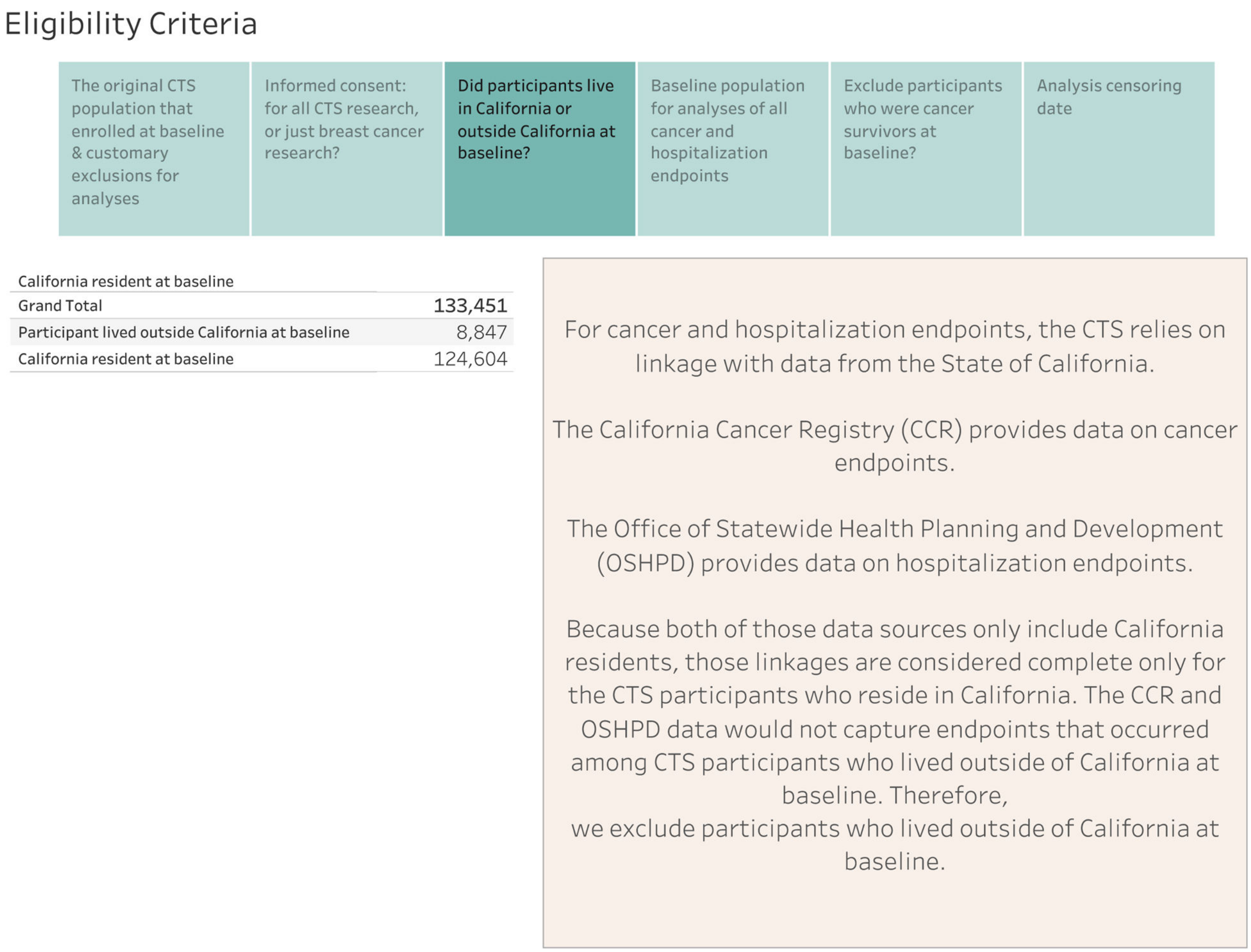

A CTS Data Warehouse
CTS Datasets: Questionnaires, Cancers, Deaths, Biospecimens, etc.
Secure, Colocated Data Repository in SDSC's Sherlock Cloud
Extract, Transform & Load
Reference Data
3rd Normal Form
Reference Data
Process Adjustment, Transformation, and Master Data Management Logic to Load to Homogenized Schema
CTS Data Warehouse
Secure CTS Analysis Environment
Statistical Analysis & Data Visualization Tools
Queries, Views, & Rollups
CTS team & all Internal & external Collaborators.

Figure 1. Schematic representation of the construction of the CTS DW.

To convert CTS datasets into a Data Warehouse, the CTS data were collocated into the secure San Diego Supercomputer Center (SDSC) environment. There, the Sherlock team normalized the data into a participant-centered data model. The Data Warehouse serves as the source of data that are presented to all investigators who query, explore, or analyze CTS data.



Eligibility Criteria

The original CTS population that enrolled at baseline & customary exclusions for analyses	Informed consent: for all CTS research, or just breast cancer research?	Did participants live in California or outside California at baseline?	Baseline population for analyses of all cancer and hospitalization endpoints	Exclude participants who were cancer survivors at baseline?	Analysis censoring date

California resident at baseline

Grand Total	133,451
Participant lived outside California at baseline	8,847
California resident at baseline	124,604

For cancer and hospitalization endpoints, the CTS relies on linkage with data from the State of California.

The California Cancer Registry (CCR) provides data on cancer endpoints.

The Office of Statewide Health Planning and Development (OSHPD) provides data on hospitalization endpoints.

Because both of those data sources only include California residents, those linkages are considered complete only for the CTS participants who reside in California. The CCR and OSHPD data would not capture endpoints that occurred among CTS participants who lived outside of California at baseline. Therefore,
we exclude participants who lived outside of California at baseline.

Figure 2. Standardized Eligibility Criteria, Inclusion Criteria, and Exclusion Criteria for CTS Analyses.

Starting in the upper left, clicking on the descriptive boxes updates the numbers and text in the bottom half of the screen. These “stories” describe the detailed cohort selection steps that each analytic project undergoes when investigators specify their inclusion, exclusion, and eligibility criteria. The output above is based on the choices that investigators make that are then transferred to the SQL template that calls these data from the CTS DW and Data Marts.

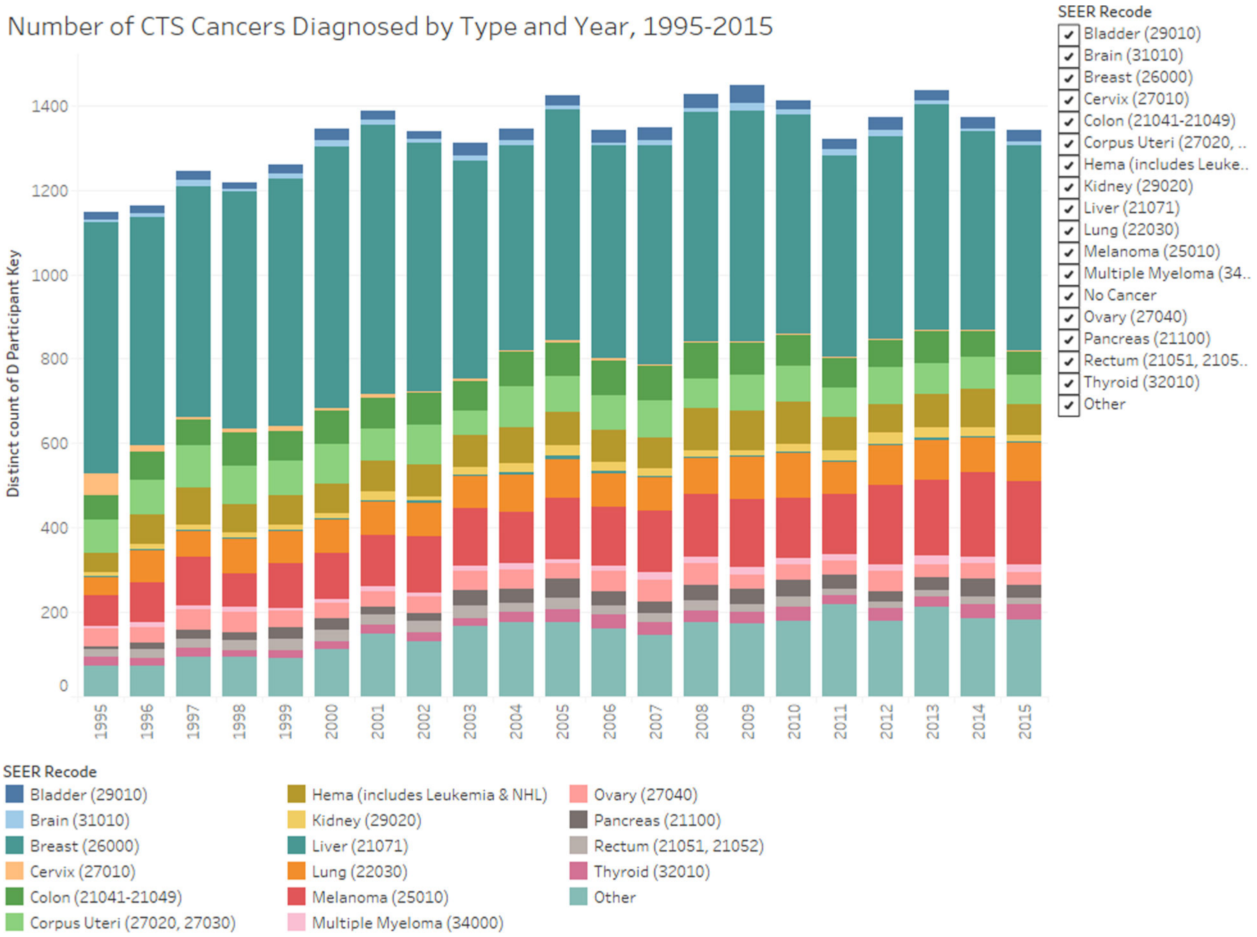
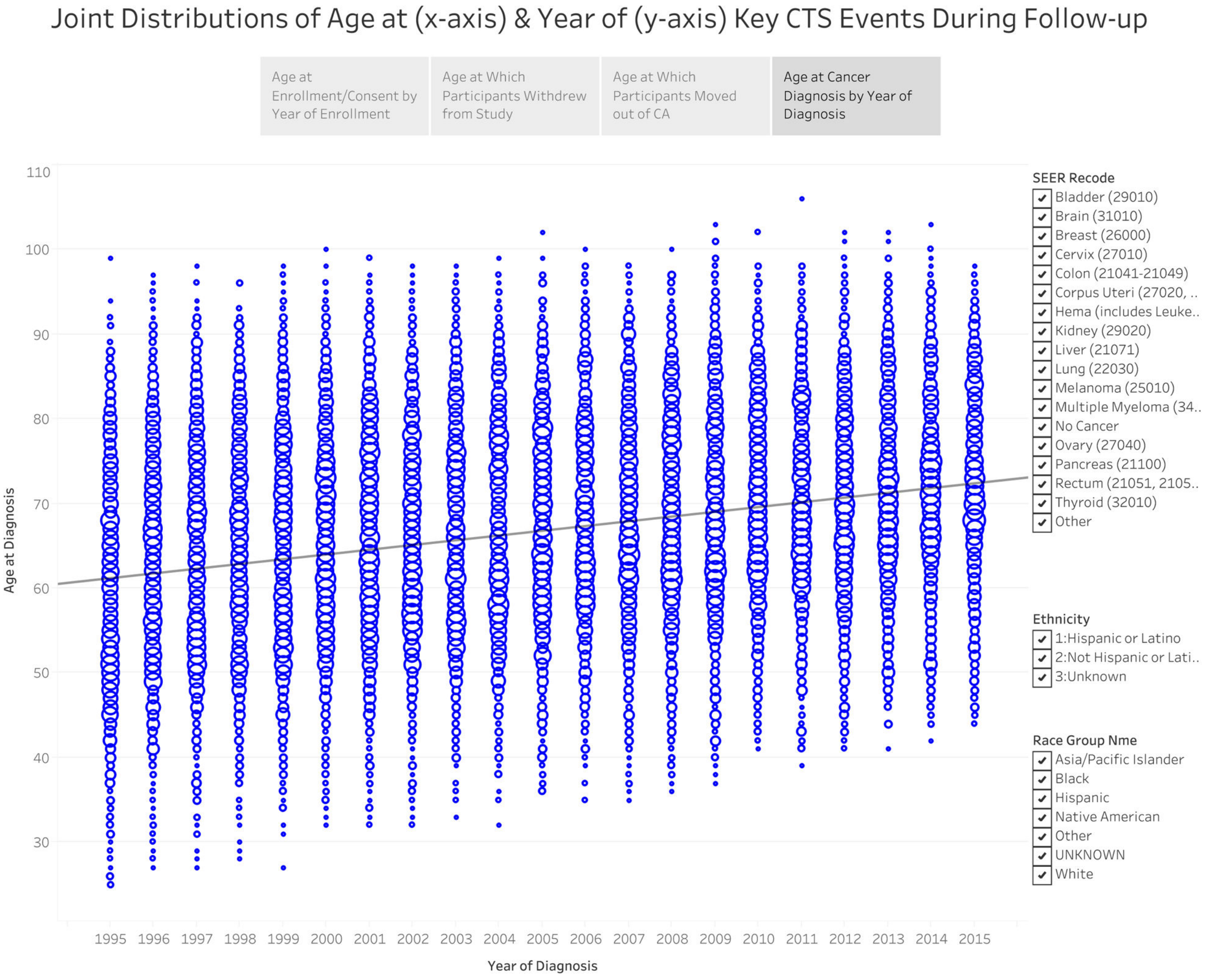


Figure 3. Numbers of Cancers Diagnosed During CTS Follow-up.

All data are interactive. Clicking on one of the SEER recode values on the right, or one of the colored boxes at the bottom, updates the figure to present data for that choice only. Right-clicking on the cells in the figure gives users the option to exclude, keep, or view those specific data, e.g., the individual data for all specific cancers diagnosed in a given site during a given year.



Joint Distributions of Age at (x-axis) & Year of (y-axis) Key CTS Events During Follow-up

Age at Enrollment/Consent by Year of Enrollment | Age at Which Participants Withdrew from Study | Age at Which Participants Moved out of CA | Age at Cancer Diagnosis by Year of Diagnosis

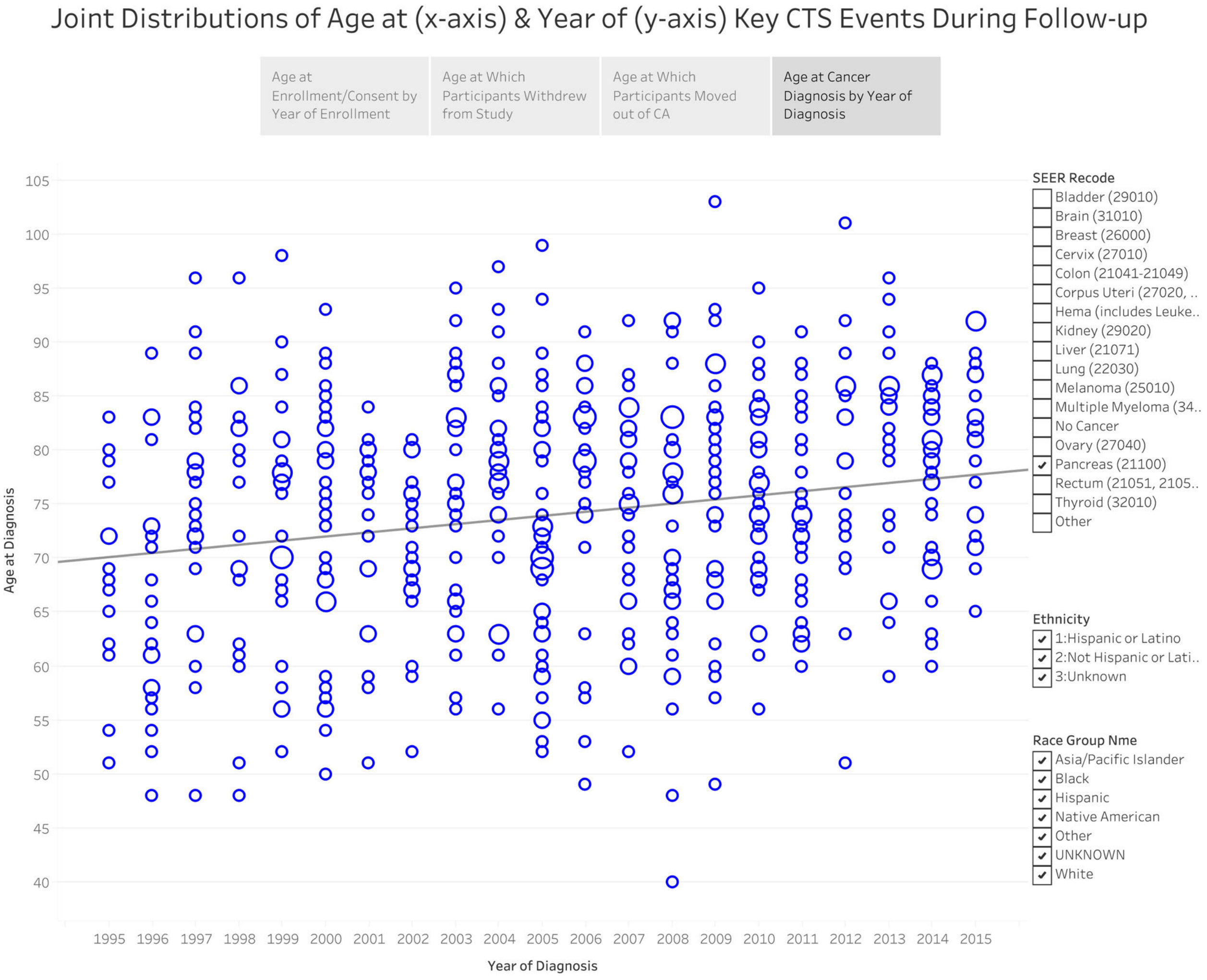
SEER Recode
☑ Bladder (29010)
☑ Brain (31010)
☑ Breast (26000)
☑ Cervix (27010)
☑ Colon (21041-21049)
☑ Corpus Uteri (27020, ..
☑ Hema (includes Leuke..
☑ Kidney (29020)
☑ Liver (21071)
☑ Lung (22030)
☑ Melanoma (25010)
☑ Multiple Myeloma (34..
☑ No Cancer
☑ Ovary (27040)
☑ Pancreas (21100)
☑ Rectum (21051, 2105..
☑ Thyroid (32010)
☑ Other

Ethnicity
☑ 1:Hispanic or Latino
☑ 2:Not Hispanic or Lati..
☑ 3:Unknown

Race Group Nme
☑ Asia/Pacific Islander
☑ Black
☑ Hispanic
☑ Native American
☑ Other
☑ UNKNOWN
☑ White

Figure 4. Scatter plot of the distribution of different age and year characteristics in the CTS.

All data are interactive. Clicking on the boxes at the top changes the figure to present different combinations, such as ages at vs. years of enrollment in the CTS. Clicking on one of the SEER recode, ethnicity, or race values on the right updates the figure to present data for that choice only. Multiple choices can be selected at the same time. Right-clicking on the data points in the figure gives users the option to exclude, keep, or view those specific data, e.g., the individual data for all specific cancers diagnosed among a specific age- and site group during a given year.



Joint Distributions of Age at (x-axis) & Year of (y-axis) Key CTS Events During Follow-up

Age at Enrollment/Consent by Year of Enrollment | Age at Which Participants Withdrew from Study | Age at Which Participants Moved out of CA | Age at Cancer Diagnosis by Year of Diagnosis

SEER Recode
☐ Bladder (29010)
☐ Brain (31010)
☐ Breast (26000)
☐ Cervix (27010)
☐ Colon (21041-21049)
☐ Corpus Uteri (27020, ..
☐ Hema (includes Leuke..
☐ Kidney (29020)
☐ Liver (21071)
☐ Lung (22030)
☐ Melanoma (25010)
☐ Multiple Myeloma (34..
☐ No Cancer
☐ Ovary (27040)
☑ Pancreas (21100)
☐ Rectum (21051, 2105..
☐ Thyroid (32010)
☐ Other

Ethnicity
☑ 1:Hispanic or Latino
☑ 2:Not Hispanic or Lati..
☑ 3:Unknown

Race Group Nme
☑ Asia/Pacific Islander
☑ Black
☑ Hispanic
☑ Native American
☑ Other
☑ UNKNOWN
☑ White

Figure 5. Scatter plot of the distribution of different age and year characteristics in the CTS: pancreatic cancer as an example.

All data are interactive. One of the SEER recode values, for pancreatic cancer, has been selected, and this automatically updates the figure to present data for participants with pancreatic cancer. Selecting another cancer site code will add those participants and present both groups on the figure. Selectin "All" will restore the figure to present all participants with cancer. Right-clicking on the data points in the figure gives users the option to exclude, keep, or view those specific data, e.g., the individual data for all specific cancers diagnosed among a specific age- and site group during a given year.

Table 1.

Essential Investigator and Staff Requirements for Cancer Epidemiology Cohort Data.

Investigators Analyzing CTS Data Must Be Able to …	Staff Managing CTS Data Must Be Able to…
Explore frequencies, counts, and summaries and exposure and outcome data	Add new data from updated linkages or new surveys
Apply inclusion & exclusion criteria for specific projects	Clean data and replace outdated values with newer or updated values
Combine and collapse different variables within analyses and projects	Maintain records of all data updates
Save and update programs, documents, and results	Ensure that users see data they're supposed to see but don't see data they're not
Share data, results, and progress	Protect the privacy and confidentiality of data
Export results for local printing	Facilitate consistent use of standards and best-practices across all projects
"Freeze," save and return to analyses, e.g., as scientific manuscripts undergo peer-review	Efficiently identify previous versions of data, for auditing or review

Table 2.

Seven Core Services Required for a Cancer Research Data Commons, and the ways in which the CTS infrastructure meets those requirements.

Core Services Required for a Data Commons …	**Key Purpose of that Service in a Research Environment**	**SDSC Sherlock Cloud Provides this for the CTS Via …**
Authentication	Identifies researchers	Active Directory & Domain Name Services
Authorization	Determines what data a user can access	3-level role-based access
Digital ID	Identify & access data	SecurID
Metadata	Assign & access metadata	Entity-Attribute-Value Model in Data Warehouse
Security & Compliance	Support controlled access	Platform meets FISMA, HIPAA & NIST CUI requirements
Data Model	Integrate data with data models	Data normalized into a 3nf w/ star-schema model
Workflow	Execute pipelines for analysis	Encourage templates & visualization